Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 1, pp. 045-055, February 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.1.045

<Application Research>

외국인 관광객 리뷰 분석을 통한 외국인 관광객 수에 영향을 끼치는 요인 분석

박홍제[1] · 이도현[1] · 김경옥[2†]

[1]서울과학기술대학교 데이터사이언스학과 / [2]서울과학기술대학교 산업공학과

Investigation on the Factors Affecting the Number of International Visitors Using Online Reviews

Hongjea Park[1] · Dohyun Lee[1] · Kyoungok Kim[2]

[1]Dept. of Data Science, Seoul National University of Science and Technology
[2]Dept. of Industrial Engineering, Seoul National University of Science and Technology

In order to attract foreign tourists, it is necessary to improve the quality of tourism services. To achieve this goal, this paper aims to extract tourists' satisfaction factors that attract foreign tourists on domestic tour spots and examine the relations between the extracted factors and the number of foreign travelers using a data-driven methodology. To do so, we collected review comments from Tripadvisor which is one of the most popular travel platforms and semantics-assisted non-negative matrix factorization is utilized to extract key satisfaction factors. Then, sentiment scores on the extracted factors are estimated using the sentiment dictionary obtained by elastic net. Unlike existing studies that usually focused on defining satisfaction factors for tourists, we investigated the effects of the factors on the number of foreign tourists using Seoul floating population data.

Keywords: Text Mining, Sentiment Analysis, Semantics-Assisted Non-Negative Matrix Factorization, Ridge Regression, Tourism

1. 서 론

관광은 '도시가 축적해 온 문화를 향유하고 공유하는 산업으로 도시경제를 견인하는 신성장동력'으로써 인적서비스 및 융·복합 산업의 비중이 높은 고부가가치 산업이며 복합 산업으로써 다른 산업의 동반성장을 유도한다(Seoul Metropolitan Government, 2018). 서울시는 이미 관광산업의 육성을 위해 2014년부터 4년 동안 서울관광발전 종합계획을 수립, 추진한 바 있다. 이를 통해 2016년 사상 최대인 1,345만 명의 외국인 관광객을 유치했고, 2018년에는 서울관광 중기 발전계획을 추진하며 2,400만 명을 목표하고 있다(Seoul Metropolitan Government, 2018). 하지만 계획안의 대부분은 관광환경 조성에 대한 내용으로 관광산업의 질적인 투자는 아직 미흡한 상황이다. 외국인 관광객의 관광지에 대한 이미지가 재방문 의도에 영향이 있다는 점을 고려하면 (임화순 등, 2017) 관광환경 조성뿐만 아니라 외국인 관광객들이 느끼는 불·만족 요인을 파악하여 관광지에 대한 이미지를 향상시키고 외국인 관광객들의 재방문율을 증가시킬 필요가 있다.

외국인 관광객들의 불·만족 요인을 파악하기 위해서 전통적으로 설문조사 기반의 연구가 진행됐다. 그렇지만, 질문에서 단어 하나 혹은 사소한 표현의 차이로도 측정의 차이가 발생할 수 있고 문항이나 답을 연구자가 직접 작성하기 때문에 연구자의 관심이나 주관이 설문지에 개입될 수 있다(Kim, 1995). 또한,

이 연구는 서울과학기술대학교 교내연구비의 지원으로 수행되었습니다.
† 연락저자 : 김경옥 교수, 01811 서울시 노원구 공릉로 232 서울과학기술대학교 산업공학과, Tel : 02-970-7286, Fax : 02-974-2849, E-mail : kyoungok.kim@seoultech.ac.kr
2020년 7월 20일 접수; 2020년 10월 10일 수정본 접수; 2020년 10월 12일 게재 확정.

설문조사는 설문 대상은 구하고 조사를 진행하는데 시간과 비용이 많이 소모된다. 특히 외국인 관광객을 대상으로 한 연구의 경우 샘플 수를 확보하는 데 어려움이 있어 연구 결과의 신뢰성에 문제를 가져다줄 수 있다.

이러한 문제점들을 개선하기 위한 방안으로 인터넷 리뷰 데이터를 활용한 분석이 주목받고 있다. 리뷰 데이터는 최근 인터넷의 발달과 스마트 폰의 보급으로 상대적으로 적은 시간 동안 충분한 양의 데이터를 확보할 수 있고, 사람들의 직접적인 의견을 잘 파악할 수 있는 분석 지표로 활용될 수 있어 객관적인 반응을 얻을 수 있다(Kim *et al*., 2015). 그뿐만 아니라 소비자들의 리뷰 수가 많이 증가함에 따라 예비 소비자들이 리뷰를 통해 정보를 얻는 온라인 구전 효과가 크게 증가하고 있어 리뷰 데이터의 중요도가 점차 증가하고 있다(Chun, 2015).

한편, 국내 관광지에 대한 인터넷 리뷰 데이터를 이용한 연구들은 주로 관광지에 대한 불 · 만족 요인을 추출하는데 국한되어 있는 경우가 많다. 그렇지만, 불 · 만족 요인과 관광지를 방문하는 관광객의 수 사이의 관계까지 알아야 보다 많은 관광객을 유치할 수 있는 방안을 효과적으로 수립할 수 있다. 기존에 관광객이 많이 찾는 지역이나 관광지에 대한 연구는 사람들의 위치를 파악할 수 있는 데이터(geo-tagging이 되어 있는 SNS 상의 글이나 사진 등)를 활용해서 사람들이 많이 방문하는 지역을 찾는 hotspot 분석 위주로 진행되었으며(Önder, Koerbitz and Hubmann-Haidvogel, 2014; Lee and Tsou, 2018), 왜 그 지역에 사람들이 많이 방문하는지를 분석한 사례를 찾기 어렵다.

이에 본 연구는 인터넷 리뷰 데이터를 기반으로 서울 주요 관광지에 대한 외국인 관광객들의 불 · 만족 요인을 추출하고 이 요인이 외국인 관광객 수에 영향을 끼치는 요인을 분석하고자 한다. 이를 위해 불 · 만족 요인 추출을 위해서는 여행 리뷰 사이트인 Tripadvisor로부터 서울 주요 관광지에 대한 리뷰를 활용해서 문장 단위로 토픽 모델링(topic modeling)을 진행한다. 그리고 특정 관광지를 방문한 외국인 관광객 수를 근사하는 자료로 서울생활인구 데이터의 단기체류 외국인 데이터를 이용하여 불 · 만족 요인에 대한 설명변수를 포함한 선형회귀 모형을 학습하여 불 · 만족 요인 중에서 외국인 관광객 수에 영향을 끼치는 요인을 분석한다.

2. 선행 연구

전통적인 외국인 관광객들의 불 · 만족 요인을 파악하기 위한 연구는 설문지법에 근거한 것이 다수이다. Jiang(2015)는 주로 중국 관광객들이 많이 찾아가는 관광지인 명동, 동대문, 인사동 등의 지역에서 중국 관광객을 대상으로 오프라인 설문을 진행해 이들의 주요 불 · 만족 요인을 밝혔다. 분석 결과 숙박 서비스, 음식 서비스, 입국 절차 등이 주요 불 · 만족 요인으로 나타났다. Choi *et al*.(2018)은 주요 3국인 중국, 일본, 미국의 관광소비자를 대상으로 한 외국인 관광객 실태조사를 활용해 연구 결과 공통적인 만족 요인은 음식과 쇼핑임을 밝혔다. 중국은 치안, 관광지매력도, 출입국 절차가 개별 만족요인이었다. 일본은 관광안내 서비스, 여행경비, 출입국절차가 개별 만족 요인이었으며 마지막으로 미국은 치안, 관광지 매력도, 숙박이 주요 만족요인으로 밝혀졌다. 하지만 이들은 직접적인 대면조사를 통해 이루어지기 때문에 시간과 비용이 많이 소모되며 충분한 샘플의 수를 확보하기 어렵고, 연구자의 주관이 설문 문항에 반영될 수 있다는 점 등의 한계점이 존재한다.

설문지법의 여러 단점으로 인해 최근에는 인터넷 리뷰를 통한 연구가 활발히 행해지고 있다. Cho *et al*.(2017)은 여행 리뷰 사이트인 Tripadvisor에서 영어로 작성된 리뷰를 수집한 뒤, 관광지를 3개의 유형으로 구분하여 토픽모델링 기법 중 하나인 잠재 디리클레 할당(Latent dirichlet allocation; LDA)을 활용해 관광지 유형별로 세부 불 · 만족 요인을 추출하였다. 이후 elastic net을 활용하여 감성사전을 구축하였고, LDA 기반의 가중 감성평가 방법론을 제안하여 요인별 감성점수를 계산하는 방식으로 감성 분석을 진행하여 추출된 요인들을 관광지별로 평가했다. elastic net을 활용한 감성사전 구축을 구축하는 방법은 감성 분석 중 모델 기반(Model-based) 감성 분석에 해당하며(Pang *et al*., 2002), 각 단어에 대한 회귀계수를 연속적인 값으로 계산하여 감성점수로써 활용할 수 있다는 장점이 있다(Kim *et al*., 2015). 특히 중요하지 않은 변수를 0으로 계산한다는 부분과 통계적으로 유의미하다는 점 역시 elastic net을 활용한 감성사전 구축의 장점이다. Kim *et al*.(2015)은 eastic net과 lasso, ridge를 각각 비교하여 elastic net이 감성사전 구축에 적합하다는 것을 보인 바 있다. 하지만 이러한 방식의 불 · 만족 요인을 추출하는 방법론은 연구자들이 미리 관광지를 유형별로 분류하였는데 여러 유형이 혼합된 성격을 가지고 있는 관광지도 존재할 수 있다는 점, 추출된 불 · 만족 요인이 실제 외국인 관광객 수에 얼마나 영향을 미치는지에 대한 검증이 되지 않은 점 등의 한계를 가지고 있다.

한편 Khatibi *et al*.(2019)은 소셜미디어 및 여행 리뷰 사이트에서 파생된 변수가 환경적인 변수와 더불어 관광객 수를 예측하는데 유의미함을 입증했다. 이들은 영국에 미국에 위치한 76곳의 국립공원, 잉글랜드에 위치한 27곳의 박물관과 미술관에 대해 연구를 진행하였다. 관광지 리뷰 수를 환경변수와 더불어 사용하였다. 그 결과 SVR(Support Vector Regression)을 사영하였을 때 93% 이상의 관광지에 대해 25% 이하의 MAPE 값을 얻어 주변 환경적인 변수들이 관광객 수 예측에 도움이 된다는 것을 보였다. Yang *et al*.(2015)은 구글 트렌드 데이터와 바이두의 검색 빈도수 데이터를 활용해 중국 하이난 지역의 관광객 수를 예측하는 연구를 진행했다. 구글 트렌드 데이터는 2010년 이후로, 바이두 검색 빈도수 데이터는 2006년 이후 데이터를 이용하였으며, 이들은 각각 9.86%와 3.11%의 MAPE 값을 보여 검색 빈도가 관광객 수 예측에 도움이 되는 것을 확인했다.

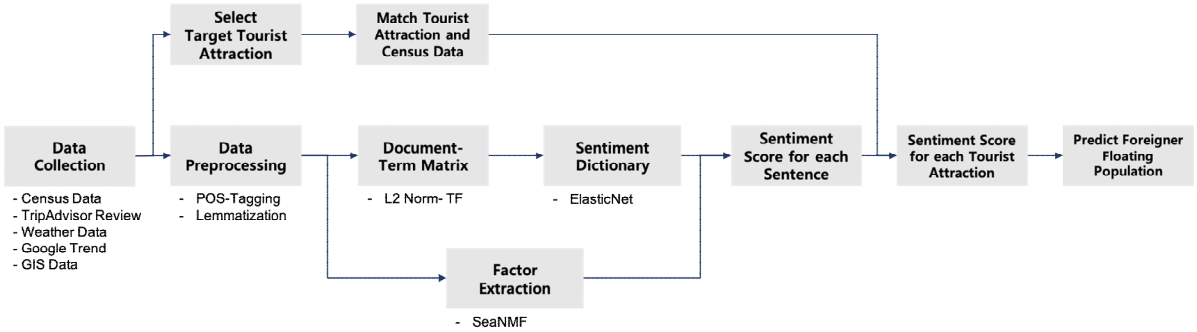


Figure 1. Research Framework

3. 연구 방법

본 장에서는 본 연구 과정에서 진행된 전처리와 방법론들을 순차적으로 설명한다. 본 연구의 프레임워크는 <Figure 1>과 같다. 먼저, 분석에 필요한 서울생활인구 데이터와 서울에 위치한 관광지의 Tripadvisor 리뷰뿐 아니라 관광객 수에 영향을 줄 만한 날씨. 구글 트렌드 데이터, 지리적 특징을 반영할 수 있는 기타 데이터들을 수집하여 전처리를 진행하였다. 이후 리뷰 데이터로부터 감성 단어와 그 단어의 감성 점수로 구성된 감성사전을 구축하고 불・만족 요인을 추출한 뒤 집계구별 요인감성점수를 산출하였다. 마지막으로 산출된 변수들과 기타 변수들을 설명변수로 사용하여 서울생활인구 데이터의 단기체류 외국인 데이터로부터 생성한 집계구별 월별 외국인 관광객 수에 대한 종속변수를 예측하는 회귀 모형을 학습하였다. 이 회귀 모형을 이용해 최종적으로 외국인 관광객 수에 영향을 끼치는 주요 요인을 추출하였다.

3.1 데이터 수집 및 전처리

1) 리뷰 데이터 : 본 연구에서는 Tripadvisor 사이트에 등재되어 있는 서울 관광지 정보 중 'Top Attractions in Seoul'을 통해 Tripadvisor에서 자체적으로 매긴 순위 순서로 300개 관광지에 대해 2019년 5월 31일까지 작성된 61,328개의 리뷰를 수집하였다. 수집된 데이터는 작성자 ID, 작성 시기, 리뷰 제목, 리뷰 내용, 평점으로 구성되어 있으며 평점은 5점 척도로 구성되어 있다. 서울생활인구 데이터가 제공되기 시작한 17년 1월 1일을 기준으로 이전 시점에 작성된 영어 리뷰의 수는 최소 5개, 이후 시점에 작성된 영어 리뷰의 수는 최소 10개 이상인 114개의 관광지를 연구대상으로 선정하였다. 연구대상에 해당하는 리뷰는 52,398개다. 리뷰 데이터는 다음과 같은 전처리를 통해 정제하였다. 먼저, 알파벳을 소문자로 변환시키고 구두점, 숫자 등 불필요한 특수문자와 불용어를 제거하였다. 이후 형태소 분석(Part-of-Speech Tagging; PoS Tagging)을 통해 명사, 동사, 형용사, 부사에 해당하는 품사만 추출하고 단어 대표형 변환기법 중 하나인 용언 추출(lemmatization)을 진행하여 다른 형태로 표현된 단어도 동일한 단어로 간주하였다.

2) 서울생활인구 데이터 : 서울생활인구는 서울 전역 6,000여 개의 기지국에서 잡힌 KT 휴대폰 신호를 바탕으로 특정한 날 특정한 시각에 특정 공간 단위(집계구, 2016년 기준)에 존재하는 모든 인구를 추정한 데이터다. 거주 인구만 파악 가능한 인구주택총조사와는 달리 시간 단위로 유동 인구수를 파악할 수 있다는 장점이 있다. 생활인구가 집계되는 공간 단위인 집계구는 통계청에서 통계정보를 제공하기 위해 구축한 최소한의 통계구역 단위이다. 서울생활인구 데이터 중에서 단기체류 외국인 데이터는 6개월 미만을 체류한 외국인에 대한 데이터이다. 한국관광공사에서 제공하는 월별 목적별 외국인 입국자에 대한 통계를 보면 2017년~2019년에 전체 외국인 입국자 중에서 약 78% 가량이 관광을 목적으로 입국했다. 이 통계를 바탕으로 해서 본 연구는 단기체류 외국인이 대부분 관광객일 것이라고 가정하고 해당 데이터를 추정 방문객 수로써 사용하기로 하고 집계구별로 총 생활인구수를 한 달 단위로 합산하였다. 다음으로, 전체 집계구 중에서 연구 목적에 부합하는 집계구를 관광지 위치, 집계구당 월 단위 서울생활인구 수를 기반으로 해서 선정하였다. 한강, 북한산국립공원 등은 수십여 개의 집계구에 걸쳐있지만 관광객이 몰리는 구역은 극히 일부에 불과하였기 때문에 관광객이 적은 집계구는 제거하였다. 그리고 각 집계구당 월 단위 서울생활인구 합이 최소 2.5만 이상이 되는 집계구들이 주요 관광지를 포함하고 있는 것을 확인하여, 이러한 조건을 만족하는 101개의 집계구를 최종 연구 대상 집계구로 선정했다.

3) 구글 트렌드 데이터 : 관광지 방문객 수는 단순 불・만족 요인뿐만 아니라 관광지의 인지도에도 크게 좌우된다(Yang, 2015). 따라서 관광지에 대한 인지도를 확인할 수 있는 데이터가 있다면 이를 반영하는 것이 방문객 수를 예측하는 데 있어서 중요한 요소가 될 수 있다. 본 연구는 관광지에 대한 인지도에 지표를 구글 트렌드를 활용하였다. 구글 트렌드란 인터넷 사용자들이 구글 검색을 통해서 검색한 내역을 각종 통계 지표로 만들어서 사용자에게 제공하는 서비스다. 트렌드 값이 높을수록 많은 사람들이 해당 키워드 검색을 많이 했다는 것을 의미한다. 구글이 외국인 관광객들이 주로 이용하는 검색엔진임을 미루어 봤을 때, 이를 효과적으로 측정할 수 있다고 판단하였다.

관광지명에 대한 변수뿐 아니라 관광객들이 선호하는 관광지 유형도 반영하기 위해 관광지 유형별 구글 트렌드 변수도 생성하였다. 동일한 명칭이 외국에도 존재하는 경우 korea을 덧붙여 검색하였다. 그리고 특정 관광지가 여러 명칭으로 불리는 경우 각각의 트렌드 값을 합산했다. 관광지 유형은 Tripadvisor에서 제공하는 관광지 유형을 활용했다. 관광지 유형을 검색키워드로 사용할 때는 해당 관광지 유형 앞에 'korea'를 덧붙여 검색량을 제한시켰다.

추가로, 구글 트렌드는 특정 기간 안에 검색된 키워드 검색량 중 최대치를 100으로 하여 나머지 키워드들을 이에 대한 상대적인 값을 제공하기 때문에 서로 다른 키워드에 대해서 검색량을 절대적으로 비교하기 어렵다. 그렇지만 한 번에 최대 5개의 키워드를 검색할 수 있어 본 연구에서는 검색량을 상대적으로 비교할 수 있게 기준 키워드를 선정하여 이들과 대상 키워드에 대해 동시에 구글 트렌드 값을 추출하였다. 최대 트렌드 값을 가지는 키워드로 롯데월드, 최소 트렌드 값을 가지는 키워드로 특수문자조합을 기준 키워드로 선정하였다.

구글 트렌드 변수는 [0, 100] 범위의 값을 갖지만 이를 [0, 1] 범위로 조정하여 사용하였다.

4) 날씨 데이터 : 날씨 데이터의 경우 기상청 국가기후데이터센터에서 제공하는 종관기상관측 데이터를 활용하였다. 종관기상관측이란 종관규모의 날씨를 파악하기 위하여 정해진 시각에 모든 관측소에서 같은 시각에 실시하는 지상관측을 말한다. 변수로 사용한 정보는 월별 평균 강수량(cm), 월별 평균 상대습도, 월별 강수일, 그리고 월별 평균기온을 사용했다. 평균기온의 경우 -0℃ 미만, 0~10℃, 10~ 20℃, 20~30℃, 30℃ 이상으로 총 5개 범주로 나눈 다음, 월별로 개별 범주에 해당하는 날이 며칠이 되는지를 변수로 만들었다.

5) 지리정보 데이터 : 지리 데이터의 경우 관광지 주변에 대중교통 혹은 인접 관광지와의 접근성이 좋을수록 많은 수의 관광객들이 방문할 수 있다는 가정 하에, 서울열린데이터광장에서 제공하는 서울시 버스 정류소 좌표 데이터(2019년 기준)와 서울시 역코드로 지하철역 위치조회 데이터(2018년 기준)를 이용하여 집계구 중심좌표로부터 주변 1km 반경 내의 주변 관광지 수, 지하철역 수, 최인접 지하철역 거리(km), 버스 정류소 수, 최인접 버스 정류소 거리(km)를 계산하여 각각을 변수로 활용했다.

3.2 정규화 회귀분석(Regularized Regression)을 통한 감성사전 구축

불·만족 요인을 추출하기 위한 기본 단어셋을 구성하고 관광지별로 불·만족 요인별 점수를 산출하기 위해 본 연구에서는 감성사전($Dict$)을 구축하였다. 감성사전을 구축에는 정규화 회귀분석을 이용했다. 회귀모델은 새로운 관측값에 대한 예측력과 모델의 해석을 위해 사용되는 경우가 일반적이다(Zou and Hastie, 2005). 그러나 설명 변수들의 강한 선형 관계로 인해 야기되는 다중공선성 문제나 모델의 과적합 문제로 인해 모델이 불안정해지며 회귀 계수가 지나치게 커지는 문제가 발생하는 경우 모델을 통한 해석이나 예측에 어려움이 생길 수 있다는 단점이 있다. 이러한 문제를 방지하기 위해 정규화 회귀분석은 목적함수에서 회귀 계수에 가중치를 주는 방식을 통해 회귀 계수의 크기를 제한한다.

리뷰를 구성하는 단어들로부터 리뷰의 평점을 잘 예측하는 모델을 통해 모든 단어들에 대한 감성점수를 획득하여 감성사전을 구축하기 위해서 본 연구에서는 정규화 회귀분석 중에서 elastic net을 이용하였다. elastic net의 목적함수는 식 (1)과 같은데 일반 선형 회귀의 목적함수($\frac{1}{2n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$)에 회귀 계수의 절댓값에 제약($l_1$ 제약)을 주는 lasso 방식과 회귀 계수의 제곱합에 제약(l_2 제약)을 주는 ridge 방식의 제약 조건이 혼합되어 더해져있다. 제약 조건에는 전체 제약 조건의 강도를 조절하는 하이퍼 파라미터(hyper parameter) α와, l_1 제약의 가중치를 조절하는 하이퍼 파라미터 l이 존재하는데 이들을 설정하면 중요하지 않은 변수의 회귀 계수를 0으로 추정해 변수 선택의 기능을 제공하면서 다중공선성 문제나 과적합 문제를 방지할 수 있다.

$$J(\boldsymbol{\beta})=\frac{1}{2n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\alpha\left\{\frac{(1-l)}{2}\sum_{i=1}^{p}\beta_i^2+l\sum_{i=1}^{p}|\beta_i|\right\} \quad (1)$$

본 연구에서는 전처리를 완료한 리뷰 데이터는 리뷰별로 Bag-of-Words(BoW) 방식으로 문서-단어 행렬(A)을 구성하고 여기에 elastic net을 적용하여 리뷰 평점을 예측하는 모델을 학습해 회귀계수를 산출하였다. 이때 회귀계수의 부호는 극성의 방향을 계수의 절대값은 극성의 정도로 해석할 수 있으므로 이를 이용해 감성사전을 구축하였다(Kim *et al*., 2015).

문서-단어 행렬을 계산할 때 단어의 실제 등장 빈도를 그대로 사용하면 리뷰 길이에 따라 리뷰별로 단어의 등장 빈도의 편차가 커지는 문제가 발생할 수 있다. 이러한 문제를 해결하고자 문서-단어 행렬에서 각 문서를 구성하는 단어들의 빈도(Term-Frequency; TF)를 식 (1)과 같이 l_2 방식으로 정규화하여 단어-빈도 행렬을 구성하였다. 식 (2)에서 $f_{t,d}$는 d번째 문서에서 단어 t에 대한 빈도이고 T는 전체 단어의 수이다.

$$f'_{t,d}=\frac{f_{t,d}^2}{\sqrt{\sum_{t}^{T}f_{t,d}^2}}. \quad (2)$$

또한, 이 모델의 목적은 리뷰를 구성하는 단어들로부터 리뷰의 평점을 잘 예측하는 모델을 통해 모든 단어들에 대한 감성점수를 획득하여 감성사전을 구축하는 것이므로 특정 문서에서만 등장하는 단어들의 중요도를 키워주는 역 문서 빈도(Inverse-Document-Frequency; IDF)를 통한 가중치 계산은 반영시키지 않았다.

감성사전 구축을 위한 elastic net의 하이퍼 파라미터는 10-fold 교차학습을 이용하여 평균제곱오차(Mean Squared Error; MSE)를 기준으로 결정하였다.

3.3 토픽 모델링을 통한 불・만족 요인 추출

본 연구에서는 한 리뷰에서 서로 다른 불・만족 요인에 대한 기술이 혼재되어 있고 각각의 불・만족 요인별로 각 요인을 나타나는 단어가 다르며 같은 요인을 기술하는 단어들은 서로 인접해서 등장할 것이라는 가정 하에 토픽 모델링을 이용해서 불・만족 요인을 정의하였다. 리뷰 단위로는 여러 불・만족 요인이 등장할 수 있지만, 문장 수준에서는 한 불・만족 요인에 대해서 기술할 가능성이 높으므로 리뷰를 문장 단위로 나눈 다음에 토픽 모델링을 진행하였다.

대표적으로 토픽 모델링에는 LDA나 non-negative matrix factorization(NMF)가 많이 사용된다. 이 중 NMF는 음이 아닌 값으로 구성된 행렬을 두 종류의 양의 행렬의 곱으로 분해(factorizing)하는 기법이다(Lee and Seung, 1999). 토픽 모델링에서는 문서-단어 행렬, A는 NMF에 의해 $A \approx WH^T$의 분해로 근사하게 된다(Shahnaz *et al*., 2006). 이 때 W와 H의 크기는 각각 $N \times K$, $T \times K$가 되는데 N, T, K는 문서, 단어, 토픽의 개수를 의미한다. 즉, NMF는 A를 각 문서를 서로 다른 토픽에 대한 가중치를 담고 있는 행렬 W와 단어를 K차원의 토픽 공간상으로 사상하는 행렬 H로 분해한다.

NMF는 문서-단어 행렬이 담고 있는 문서 혹은 문장에서 단어들의 동시 출현 빈도에 기반하고 있는데, 길이가 짧은 문서 내에는 의미 있는 단어의 수가 적어서 동시 출현 빈도만으로는 의미 있는 토픽을 추출해 내기 어렵다. 그러므로 문장 단위로 토픽 모델링을 진행할 때에는 기존의 방법으로는 제대로 불・만족 요인을 추출하기 어려울 것이라 판단하였다. 이에 본 연구에서는 짧은 텍스트에 특화된 토픽 모델링 알고리즘인 semantics-assisted non-negative matrix factorization(SeaNMF)를 활용하였다.

SeaNMF는 길이가 짧은 문서를 이용해 토픽 모델링을 수행하기 위해 NMF에 단어의 시맨틱 정보를 활용한 알고리즘이다(Shi *et al*., 2018). 이 알고리즘은 단어의 시맨틱 정보를 NFM에 반영하기 위해 단어 상관 행렬을 사용하였다. 이와 같은 방식은 워드 임베딩에서 널리 쓰이는 negative sampling을 이용한 skip-gram 방법이 특정 단어와 주변 단어 사이의 상호정보량(pointwise mutual information, PMI)에 기반한 단어 상관 행렬을 분해하는 것과 동일하다는 것을 보인 Levy and Goldberg(2014) 연구를 바탕으로 한 것이다.

본 연구에서는 리뷰 전처리 과정 중 용언 추출 작업까지만 진행한 뒤 불용어 등을 제거하고 리뷰를 문장 단위로 나누었다. 그리고 단어의 수가 5개 미만인 문장은 제외하고 얻어진 문장-단어 행렬(A)을 SeaNMF를 이용해 분해하였다. SeaNMF의 결과로 얻어진 단어-토픽 행렬(H)로부터 문장-토픽 행렬(W) 토픽별로 가중치가 높은 단어들을 추출하여 불・만족 요인을 정의하였다. 토픽을 기반으로 불・만족 요인을 정의하는 데는 기존의 설문조사 기반의 관광지에 대한 불・만족 요인에 대한 연구(Yeum and Lee, 2015; Yu, Lee and Lee, 2016; Lim and Nam, 2017)를 참고하였다. 그리고 SeaNMF로부터 얻어진 문장-토픽 행렬(W)은 관광지별 요인별 감성 점수를 산출할 때 이용하였다.

3.4 관광지별 요인별 감성 점수 산출

관광지별 요인별 감성 점수를 산출하기 위해서 먼저 리뷰별로 요인별 감성 점수를 산출하는 작업을 수행하였다. 한 리뷰 안에서는 여러 가지 불・만족 요인에 대한 언급이 존재할 수 있다. 그러므로 요인별 감성 점수를 산출하기 위해 각 문장 단위로 그 문장이 어떤 요인에 대한 정보를 담고 있는지 문장별 요인확률을 구하였다.

문장의 요인확률 계산에는 3.3절에서 구한 W를 이용하였다. SeaNMF는 NMF에 기반하므로 W를 구성하는 모든 값은 0 또는 양수이다. 각 문장별로 토픽의 비중을 구하기 위해서 토픽의 가중치 합이 1이 되게끔 W를 정규화하였다. 문장의 감성 점수는 정규화한 TF 벡터, $x'_s = [f'_{t,s}]$의 $f'_{t,s}$ 값과 t의 감성 점수를 곱해서 다 더한 값으로 정의하였다. 그리고 그 문장의 요인별 감성 점수는 해당 문장의 요인별 비중과 문장의 감성 점수를 곱해서 사용하였다. 이를 통해서 특정 문장에서 특정 요인의 가중치가 크다면 해당 요인의 평균 감성 점수에 이 문장의 감성 점수가 더 크게 반영될 수 있게 했다. 문장 단위의 요인별 감성 점수가 구해지면, 개별 리뷰에 등장하는 모든 문장을 활용해서 리뷰/요인별로 감성 점수의 가중 평균을 계산하였다. 마찬가지로 관광지/요인별 감성 점수는 해당 관광지에 대한 리뷰의 요인별 감성 점수를 평균하여 계산하였다. 이 때 월별로 다른 감성 점수를 계산하기 위해서 해당 월 이전의 리뷰만을 사용하여 특정 월의 요인별 감성 점수를 계산하였다.

3.5 외국인 관광객 수에 영향을 끼치는 요인 분석

외국인 관광객 수에 영향을 끼치는 관광객의 불・만족 요인을 추출하기 위해서 ridge 방식의 정규화 회귀 모형을 사용하였다. 설명변수와 종속변수의 관계를 통계적으로 검증하기 위해 ridge 방식의 정규화 회귀 모형을 사용하였으며 회귀 모형의 유의성을 검증하기 위해 F-test(Bae, Kim and Kim, 2014)와 설명변수의 유의성을 검증하기 위해 t-test(Cule, E., Vineis, P. and De Iorio, M., 2011)를 진행했다.

관광지별로 외국인 관광객 수에 대한 통계자료는 구하기 어려워 3.1절에서 설명한 서울생활인구 데이터 중 단기체류 외국인 데이터를 이용했다. 단기체류 외국인의 대부분이 관광객일 것이라는 가정하에 이 데이터로 구한 월별 집계구별 단기체류 외국인 수로부터 종속변수를 구하였다. 집계구별로 단기체류 외국인 수의 편차가 매우 크고 일부 집계구에서 다른 집계구에 비해 단기체류 외국인 수가 매우 많아서 월별 집계구별 단기체류 외국인 수에 $\log_{10}$ 을 취한 값을 최종 종속변수로 하였다.

설명변수로는 먼저 관광지별 요인별 감성 점수 변수를 이용하였다. 일부 집계구는 둘 이상의 관광지를 포함하고 있어 해당 집계구에 포함된 모든 관광지의 요인별 감성 점수를 요인별로 평균하였다. 집계구별로 특정 월의 요인별 감성 점수를 구한 다음에 이로부터 파생변수를 생성하였다. 생성된 감성 점수 변수를 살펴본 결과 이들 사이의 상관계수는 쇼핑과 먹거리를 제외하고는 크지 않았지만 안내서비스를 제외하고는

나머지 요인은 감성 점수가 음의 값을 갖는 샘플이 30%를 넘지 않았고, 경험, 교통/이동수단, 자연경관/산책로, 먹거리는 10%도 채 되지 않았다. 이는 긍정 리뷰가 부정 리뷰에 비해 더 많기 때인데 이로 인해 감성 점수 변수들의 종속변수와의 상관계수는 박물관/전시, 자연/공원을 제외하고는 모두 음의 값을 가진다. 즉, 대부분 요인에서 감성 점수가 높으면 외국인 관광객이 감소한다는 뜻이므로 이를 바탕으로 외국인 관광객 수에 더 큰 영향을 끼치는 요인을 판별하기 어려워 이들을 두 가지 파생변수를 생성하였다.

첫 번째 파생변수는 감성 점수가 부정(음수)인지 여부를 가리키는 이진 변수이다. 요인의 감성 점수가 부정인 경우가 많지 않으므로 부정인 거 자체가 외국인 관광객의 방문에 영향을 끼칠 수가 있을 것이라고 생각해 이 변수를 생성하였다. 두 번째 파생변수는 요인별 감성 점수의 순위와 요인별 가중치를 동시에 고려한 변수이다. 각 샘플별로 감성 점수에 따라 요인의 순위를 정하고 이 값의 역수를 취한 다음에 요인의 가중치를 곱해서 변수를 생성하였다. 요인별 감성 점수가 모두 긍정적이라고 할지라도 그 안에서 순위는 다르므로 어떤 요소가 다른 요소보다 더 나은지에 따라 관광객 수가 달라질 수 있을 수 있어 이와 같은 변수를 생성하였다. 여기에 요인의 가중치를 곱한 것은 보다 많이 언급된 요인의 순위가 높은 것에 더 큰 가중치를 주기 위함이다.

추가로 3.1절에서 기술한 것처럼 요인별 감성 점수 변수 외에도 구글 트렌드, 날씨, 지리정보로부터 얻은 변수도 설명변수로 활용했으며 광화문 지역은 다른 지역에 비해 관광객 수가 10배 이상 많아서 이들 집계구를 나타내는 지시변수도 추가하였다.

다중공선성 효과를 완화할 수 있는 ridge 방식의 정규화 회귀를 사용하지만 온도 변수의 경우 월별 일수가 28, 30, 31 중 하나이기 때문에 나머지 한 변수의 값은 다른 변수의 값이 정해지면 거의 정확하게 예측이 가능하다. 이에 '< 0' 변수는 제거하였다. 마찬가지로 감성 점수 변수 중 쇼핑과 먹거리를 제외하고는 두 파생 변수를 포함해서 각 그룹 내에서는 VIF가 다 10 이내로 다중공선성이 크지 않지만, 이 세 그룹의 변수를 모두 사용할 때는 다중공선성이 크게 증가하고 변수 수가 과도하게 많아져 특정 요인의 영향력을 해석하기에 어려움이 따를 소지가 크므로 이들 중 하나 또는 두 가지 종류의 변수들만 선택하여 서로 다른 학습 데이터를 구성하였다. 최적의 조합은 10-fold 교차검증을 통해 adjusted R-square를 이용해 결정하였다. 최적 조합이 결정된 이후에 전체 데이터를 이용해서 회귀 모델을 학습했다.

4. 실험결과

4.1 감성사전 구축

10-fold 교차검증을 통해 구한 최적의 하이퍼 파라미터는 α,l는 각각 e^{-11}, 0.5였다. 이 때 10-fold 교차검증으로 구해진 MSE는 0.47이었다. 그리고 감성 점수가 4점 이상이면 긍정, 2점 이하이면 부정, 나머지는 중립인 세 범주로 나눴을 때 예측된 값이 실제 범주를 얼마나 잘 예측하는지 10-fold 교차검증으로 정확도를 측정한 결과 0.86이었다. 이 정도 성능은 감성사전을 구축하는데 큰 문제가 없을 것이라고 판단하여 최적의 하이퍼 파라미터를 이용해 전체 데이터를 다시 학습시켜 최종모델의 회귀계수를 단어의 감성점수로 활용하여 감성사전을 구축했다.

<Table 1>은 최종적으로 얻어진 감성사전에서 감성점수의 크기가 상위 10위 안에 드는 긍정어와 부정어를 보여주고 있다. <Table 1>을 보게 되면 긍정단어의 감성점수의 크기가 부정단어의 감성점수의 크기보다 작은 걸 볼 수 있는데 이는 감성사전 구축에 사용한 전체 리뷰 중 평점인 5점과 4점인 각각 48.63%, 36.87%로 85%에 육박하는 리뷰가 긍정 리뷰이기 때문이다. 긍정어와 부정어를 살펴보면 일반적으로 사용할 수 있는 단어도 있지만, "yummy", "unfriendly", "overpriced", "dirty"와 같이 관광지의 특정 요인에 대해 만족 혹은 불만족을 표현할 수 있는 단어도 존재한다.

Table 1. Top 10 Positive and Negative Words

Positive Word	Sentiment Score	Negative Word	Sentiment Score
automatically	1.3906	waste	-4.3843
donation	1.3347	unfriendly	-4.1064
favorite	1.3135	overrated	-3.4291
excellent	1.2643	underwhelming	-3.1048
incredible	1.2403	boring	-3.0519
dream	1.2217	poor	-3.0173
stunning	1.1302	disappoint	-2.8735
yummy	1.1208	terrible	-2.8253
amazing	1.1180	overpriced	-2.7941
brilliant	1.1171	dirty	-2.6079

4.2 불 · 만족 요인 추출

최대한 서로 구별이 되는 토픽을 추출하기 위해서 토픽의 개수를 바꿔가며 SeaNMF의 하이퍼 파라미터α와 β를 조정하여 여러 번의 실험을 진행하였다. α는 스케일, β는 W에 대한 l_1 제약의 크기를 조절하는 역할을 한다. 토픽의 개수는 10~20, α는 0~1, β는 0~2 사이의 값을 바꿔가며 모델 결과를 해석했다. 토픽의 개수는 적을수록 토픽 내에 여러 요인들이 동시에 등장하는 경우가 많아 토픽을 요인으로 정의하기 어려웠다. 반대로 토픽의 개수가 많을수록 중복되는 내용의 토픽이 다수 등장하는 문제가 있었다. α의 경우 값이 작을수록 일반적으로 동시에 등장하는 단어들로 이루어진 토픽들을 얻을 수 있었으며 이 경우 먹거리, 쇼핑 등에 대한 토픽들을 얻을 수 없었다. 반면 α가 커질수록 토픽 내에 복합명사를 이루거나 연관성이 높은 명사들로 이루어진 토픽들을 얻을 수 있었으나 α가 너무 크면 복합명사로 이루어진 토픽들의 수가 너무 많아져 일반적으로 동시에 등장하는 단어들로 이루어진 토픽의 수가 줄어드는

문제가 있었다. β의 경우 값이 작을수록 여러 토픽에서 동시에 등장하는 단어들의 수가 많아져 토픽을 정의하는데에 있어 어려움이 있었으며, β가 커질수록 이러한 경향이 줄어들어 토픽을 보다 쉽게 정의할 수 있었다. SeaNMF의 실험 결과 토픽의 개수는 15, α와 β는 0.33, 2.0로 설정했을 때 토픽 해석이 가장 용이하다고 판단되어 최종적인 모델을 선택하였으며, 최종 모델의 결과를 기준으로 불·만족 요인을 정의하였다.

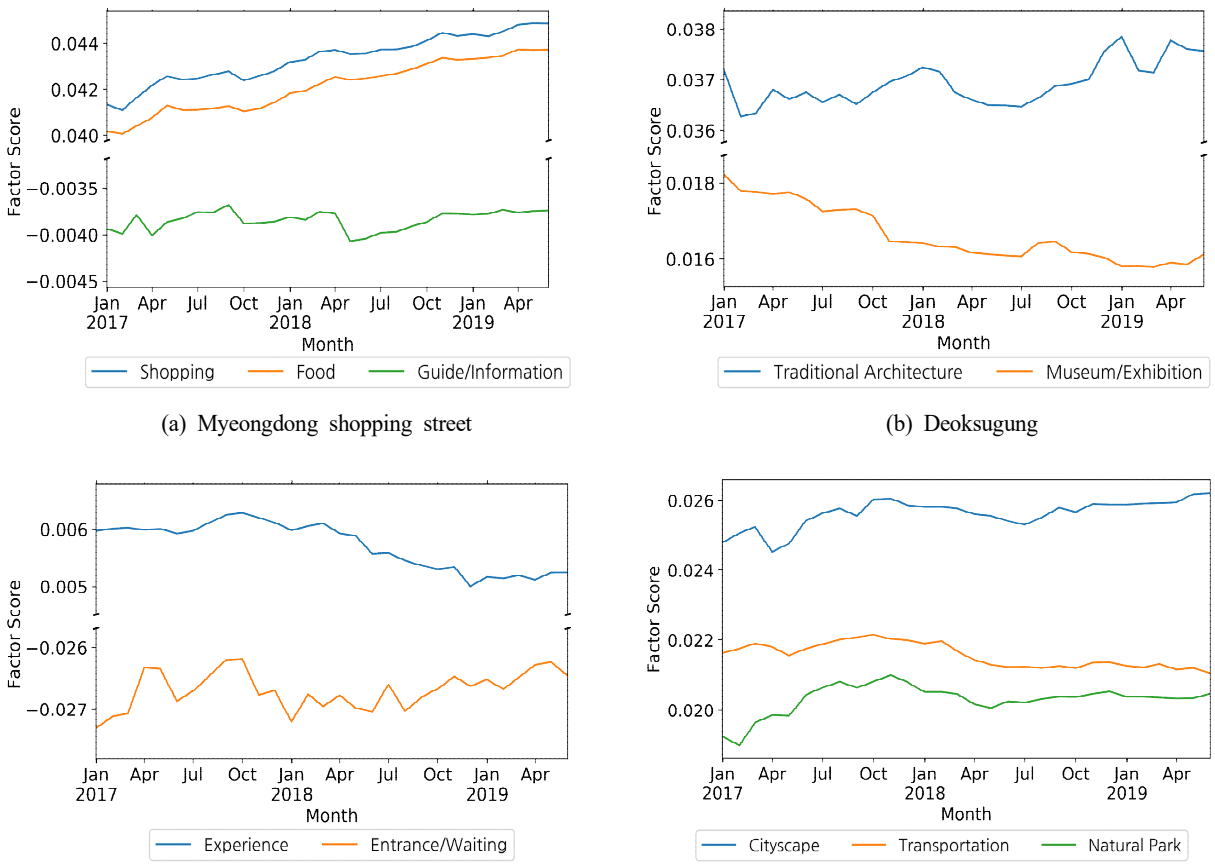
불·만족 요인은 가중치가 높은 단어를 뽑아서 정의하는데, 일부 토픽들은 같은 요인으로 묶을 수 있어 최종적으로는 총 10개의 불·만족 요인을 추출하였다. 이렇게 정의한 10개의 불·만족 요인과 이들 토픽을 정의하는데 활용한 단어들은 <Table 2>와 같다.

4.3 관광지별 요인별 감성 점수

3.3절에서 기술한 방법을 따라 4.2절에서 정의된 10개의 요인에 대해 월별 관광지별 감성 점수를 계산하였다. 연구대상 관광지 중에서 리뷰 수가 많으면서 요인별로 감성의 점수의 차이가 관광지의 특징을 잘 보여줄 수 있는 일부 관광지(명동 쇼핑거리, 덕수궁, 롯데월드, 청계천)를 선정해서 선택된 요인에 대해서 감성 점수의 변화를 <Figure 2>에 나타냈다.

Table 2. Extracted Satisfaction Factors

Factor	Keyword
Experience	visit, night, trip, free, experience, attraction, morning, tourist, site, evening
Entrance/ Waiting	time, spend, hour, night, wait, long, lot, look, people, ticket
Museum/ Exhibition	korean, traditional, history, museum, culture, house, exhibit, memorial, display, dress
Cityscape	area, city, view, beautiful, night, building, look, tourist, amazing, light
Traditional Architecture	palace, garden, building, museum, guide, royal, entrance, gate, history, national
Transportation	walk, station, subway, metro, stair, bus, train, card, path, distance
Guide/ Information	staff, problem, ask, issue, online, counter, service, helpful, help, reservation
Natural Park	tree, stream, flower, water, pond, bridge, river, plant, light, forest
Shopping	store, sell, clothe, accessory, product, souvenir, cosmetic, jewelry, item, brand
Food	fry, cake, restaurant, rice, grill, potato, egg, coffee, cheese, chicken



(a) Myeongdong shopping street

(b) Deoksugung

(c) Lotte world

(d) Cheonggyecheon stream

Figure 2. Monthly Sentiment Scores of the Selected Factors

명동쇼핑거리에서 감성 점수가 가장 높은 요인은 쇼핑과 먹거리이고 감성 점수가 낮은 요인은 안내 서비스이다. 쇼핑과 먹거리에 대한 점수는 꾸준히 상승하고 있는 반면, 안내 서비스의 경우 큰 변화 없이 낮은 점수를 보인다. 이는 다양한 상점, 식당, 길거리 음식 등 서비스업이 주인 길거리 관광지이다 보니 체계적인 관광 안내가 제한적이어서 그런 것으로 보인다.

덕수궁은 10개 요인 중에서 관광지의 특성과 관련이 깊은 전통건축과 박물관/전시 요인에서 꾸준히 긍정적인 평가를 받고 있다. 그렇지만 상승세를 보이는 전통건축과는 달리 박물관/전시의 감성 점수는 하락세를 보이고 있어 대조적이다.

다른 관광지와 달리 롯데월드에서 특징적으로 나타나는 요인은 경험과 입장/대기 요인이다. 경험에 대한 감성 점수는 계속 양수이지만 2018년 이후로 감소세에 있으며, 입장/대기 요인은 계속 음수이지만 조금씩 증가하고 있다. 사람이 많이 방문하는 시기에는 인기 놀이기구는 긴 대기시간을 감수해야 하는데, 롯데월드의 경우에는 롯데월드 앱에서 인당 3회까지는 예약을 통해 빠르게 입장할 수 있게 하고 있다. 출시는 2015년이지만 리뷰를 살펴보면 외국인 관광객들은 최근 들어서야 이용하기 시작했고, 이것이 입장/대기의 감성 점수 향상에 영향을 준 것으로 보인다.

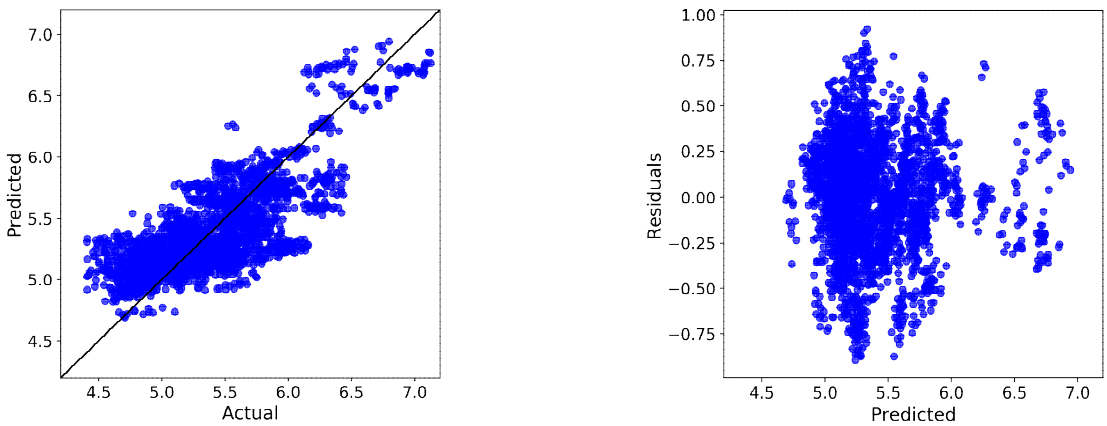
마지막으로 청계천 도시경관, 교통/이동, 자연/공원에서 높은 점수를 보인다. 빌딩 숲 사이에 있는 도심천인 데다가 걷기 좋게 조성된 산책로로 인해 도시경관과 자연/공원 양쪽 모두에서 긍정적인 점수를 받을 것으로 보이며 주변에 도보로 이동 가능한 지하철역이 많은 것이 교통/이동에서 긍정적인 리뷰를 받은 이유로 보인다.

4.4 외국인 관광객 수에 영향을 끼치는 요인 분석

외국인 관광객 수에 영향을 끼치는 요인을 추출하는 데는 ridge 방식의 정규화 회귀 모형을 활용하였다. 학습 데이터는 서울생활인구가 제공되는 17년 1월부터 19년 5월까지 29개월 동안 101개 집계구의 월별 단기체류 외국인 데이터를 활용하였으며 데이터 수는 총 2,929개이다. 10-fold 교차검증을 통해 최종 모델은 두 파생변수의 조합을 이용한 모델로 결정했다. 요인별 감성 점수 변수를 이용한 모델은 두 파생변수 중 한 종류만 이용한 두 모델보다도 성능이 떨어져 세 종류의 변수 중에서 가장 낮은 설명력을 보였다.

최종 모델에 대해서 F-test를 진행한 결과 p-value가 0.001보다 작아 통계적으로 유의함을 확인하였다. 그리고 모델의 성능을 평가하기 위해서 adjusted R-square 외에도 회귀에 사용하는 평가 지표인 MSE(mean squared errors)와 MAPE(mean absolute percentage error)도 계산해보았다. 그 결과 adjusted R-square, MSE, MAPE는 각각 0.61, 0.08, 0.04로 나타났다. 또한 예측값이 단기체류 외국인 수의 차이를 어느 정도로 잘 설명하는지 확인해 보기 위해서 <Figure 3>과 같이 실제 종속변수의 값과 예측값을 비교한 그래프와 잔차 그래프를 확인해보았다. 실제값이 작을 때에는 약간 크게 예측하는 경향이 있지만 전체적으로는 이 모델이 집계구별로 단기체류 외국인 수의 차이를 잘 설명한다고 볼 수 있다. 잔차도 종속변수의 값이 몰려있는 구간에서의 변동성이 다른 곳보다는 크지만 0을 기준으로 대칭 형태를 띠고 있어 정규분포를 이루므로 최종 모델을 이용해 단기체류 외국인 수에 영향을 끼치는 요인을 찾는데 활용 가능하다고 결론지었다.

회귀 모형의 회귀계수에 대해 t-test를 진행하여 통계적 유의성을 검정하고 <Table 3>으로 정리하였다. 절편을 제외하고 총 36개 변수 중 25개 변수의 계수가 유의수준 0.05 수준에서 유의한 것으로 확인됐다. 유의한 변수 중 부정 여부를 나타내는 첫 번째 파생변수에서는 입장/대기, 자연/공원(유의수준 0.1 수준에서 유의함)을 제외한 나머지 요인에 대한 변수는 유의한 것으로 나타났고, 두 번째 파생변수에서는 경험, 박물관/전시, 전통건축만 유의하지 않은 것으로 나타났다. 기타 변수 중에서는 지리적 변수는 모두 유의하게 나타났으나 날씨 변수는 20~30℃를 제외하고는 모두 유의하지 않았다. 2017~2019년 한국관광공사의 외국인 입국자 통계에 따르면 관광을 목적으로 방문하는 외국인의 수는 월평균 90만 명에 못 미치는 1~2월을 제외하고는 월평균 100만 명 수준에서 거의 차이가 나지 않는 것으로 보아 우리나라를 방문하는 외국인들에게 날씨는 그리 중요한 요소는 아닌 것으로 보인다.



(a) Comparison of linear regression model (b) Residual plot

Figure 3. Monthly Sentiment Scores of the Selected Factors

Table 3. Regression Coefficients and Their Statistics

Features			β	$se(\beta)$	t	$p-value$
	Intercept		4.4489	0.1108	40.9418	0.0000***
Satisfaction Factors	Binary for Negative or not	Experience	0.2927	0.5269	7.6190	0.0000***
		Entrance/Waiting	0.0092	0.5932	0.3730	0.7091
		Museum/Exhibition	-0.1619	0.4553	6.5530	0.0000***
		Cityscape	-0.0563	0.6189	2.1930	0.0283***
		Traditional Architecture	-0.1661	0.5178	7.5160	0.0000***
		Transportation	0.1097	0.4364	3.5320	0.0004***
		Guide/Information	-0.0727	0.5739	2.6880	0.0072***
		Natural Park	-0.0840	0.6185	1.8580	0.0632*
		Shopping	0.4284	0.5334	12.8010	0.0000***
		Food	-0.1966	0.5450	4.6510	0.0000***
	Weighted Rank	Experience	0.3653	0.4352	0.8460	0.3973
		Entrance/Waiting	-4.4795	0.5642	3.2200	0.0013***
		Museum/Exhibition	0.1691	0.4313	1.3600	0.1738
		Cityscape	0.6324	0.5492	6.1820	0.0000***
		Traditional Architecture	3.2883	0.3715	6.8660	0.0000***
		Transportation	1.3779	0.4448	4.4240	0.0000***
		Guide/Information	8.9281	0.5233	2.3970	0.0165**
		Natural Park	2.1287	0.4704	6.8850	0.0000***
		Shopping	1.9243	0.5312	5.0860	0.0000***
		Food	2.5264	0.5878	3.7950	0.0001***
Other Features	Google Trend for Category of Attraction		0.1283	0.4495	2.4160	0.0157**
	Google Trend for Attractions		0.1823	0.4935	5.5180	0.0000***
	The Number of Locally Attractions		0.1182	0.4814	10.7860	0.0000***
	The Number of Attractions in Census		0.0874	0.4261	13.5030	0.0000***
	The Number of Adjacent Bus Stops		0.2379	0.4680	9.0540	0.0000***
	The Number of Adjacent Subway Stations		0.0304	0.4631	6.6790	0.0000***
	Distance between Nearest Bus Stops		0.3428	0.4328	4.7830	0.0000***
	Distance between Nearest Subway Stations		-0.1695	0.4308	5.5290	0.0000***
	Binary for Gwanghwamun Area or Not		0.6698	0.3780	16.2810	0.0000***
	Average Daily Rainfall		-0.0862	0.6133	1.0180	0.3086
	Average Humidity		0.2443	0.6672	1.5480	0.1217
	The Number of Precipitation Days		-0.0411	0.5876	0.4940	0.6213
	0~10℃		-0.0209	0.6033	0.5240	0.6005
	10~20℃		-0.0364	0.6518	0.9450	0.3448
	20~30℃		-0.0646	0.6511	2.0000	0.0455**
	30℃ or more		0.0001	0.3719	0.0020	0.9983

*, ** and *** means statistically significant at the significance level of 0.1, 0.05 and 0.01, respectively.

불・만족 요인에 대한 첫 번째 파생변수의 계수를 살펴보면, 유의한 변수 중에서 입장/대기, 쇼핑은 종속변수와 양의 관계를 보이는데, 이는 이들 요인의 감성 점수가 부정적이면 외국인이 증가한다는 뜻으로 이는 상식과는 배치된다. 쇼핑에서 부정적인 점수가 나타나는 샘플은 전체 샘플의 10% 이내인데다 명동에 위치한 일부 쇼핑센터, 코엑스 등 쇼핑을 목적으로 많은 관광객이 방문하는 장소를 다수 포함하고 있었다. 이곳들은 외국인에게도 쇼핑으로 유명한 곳인 만큼 많은 외국인이 방문하지만 불만족한 경우도 다수 발생해 이들의 의견이 강하게 반영되었을 가능성이 있다. 특히 이들 관광지는 '혼잡하게 하다(overcrowd)'와 같은 다수 등장하는 점으로 보아 많은 인파에 의해 만족도가 떨어진다는 것을 알 수 있다.

불・만족 요인에 대한 두 번째 파생변수의 계수를 살펴보면, 유의한 변수 중에서 입장/대기는 계수가 음수이지만 나머지 요인은 양의 계수를 갖는다. 이 변수의 계수가 양수라는 것은 해당 요인이 리뷰에서 높은 비중을 보이면서 순위가 올라가게 되면 단기체류 외국인이 증가하는 데 기여한다는 걸 의미한다. 입장/대기에 대한 언급 빈도가 높은 롯데월드에서 감성 점수가

음수이면서 방문객이 많아 이 요인의 계수는 양수로 추정된 것으로 보인다. 이런 테마파크는 줄을 서는 상황이 특수한 것은 아니므로 입장/대기에 불만이 있더라도 즐길거리가 충분하다면 많은 관광객들이 방문하는 것으로 보인다.

입장/대기를 제외한 요인들의 계수의 크기를 비교해 보면, 가장 큰 것은 안내서비스, 입장/대기, 전통건축, 교통/이동, 먹거리, 자연/공원 순이었고, 도시경관은 다른 요인에 비해 계수의 크기가 작았다. 안내 서비스는 평균 순위와 가중치가 높지 않은데도 계수가 값이 큰 것으로 보아 안내 서비스가 개선되면 외국인 관광객의 방문을 효과적으로 늘릴 수 있을 것으로 기대된다. 가중치가 가장 높은 요인은 쇼핑과 먹거리인데 두 번째 파생변수에서 먹거리의 계수가 더 큰 것으로 보아 먹거리에 대한 만족도가 높은 관광지가 상당히 선호되는 걸 알 수 있다. 자연/공원은 평균 순위가 쇼핑, 먹거리 다음으로 세 번째이며 청계천, 여의도 공원, 남산 공원 등 이 요인의 순위가 3위 이상인 곳에서 평균 단기체류 외국인이 나머지 관광지에 비해 많았다. 이를 바탕으로 걷기 좋은 관광지도 외국인 관광객에게 꽤 선호되는 장소라고 결론지을 수 있다.

5. 결 론

본 연구에서는 인터넷 관광지 리뷰를 활용하여 외국인 관광객 불 · 만족 요인 추출하고 이를 바탕으로 이들이 요인과 외국인 방문객 수 간의 관계를 분석하였다. 리뷰로부터 SeaNMF를 이용해 경험, 입장/대기, 박물관/전시, 도시경관, 전통건축, 교통/이동, 안내 서비스, 총 10개의 불 · 만족 요인을 추출하였고 서울생활인구의 단기체류 외국인 데이터를 이용해 관광지가 위치한 집계구에서 단기체류 외국인의 수와 추출한 불 · 만족 요인별 감성 점수와의 관계를 파악하였다. 분석 과정에서 보다 설명력이 높은 회귀 모델을 얻기 위해 감성 점수 변수에서 감성 점수가 부정인지 여부를 나타내는 변수와 요인의 가중치를 고려한 요인 간 감성 점수의 순위 변수를 추가적으로 생성하고 서로 다른 조합의 학습 데이터로부터 여러 개의 학습 모델을 얻은 다음에 10-fold 교차검증으로 최적의 모델을 선정하였다.

분석 결과 감성 점수보다는 감성 점수가 음수인지 여부와 요인 간의 감성 점수의 가중 순위가 단기체류 외국인의 수를 더 잘 설명한다는 것을 알 수 있다. 경험, 자연/공원, 쇼핑, 먹거리 요인에 대해서는 평균적으로 평가가 부정적이더라고 단기체류 외국인이 많았으며 특히 자연/공원, 먹거리는 순위가 올랐을 때 다른 요인에 비해 단기체류 외국인 증가에 크게 영향을 끼침을 확인하였다.

하지만 본 연구는 장기적인 관광지별 외국인 방문객 수에 대한 통계를 구할 수 없어, 서울생활인구의 단기체류 외국인 데이터를 활용했다는 한계가 있다. 외국인 입국자의 90% 가까이가 관광을 목적으로 입국하기는 하지만, 단기체류 외국인 중에서는 비즈니스 목적으로 방문하는 경우도 있으며 서울생활인구 데이터를 활용하면서 집계구 단위로 분석을 진행하였기 때문에 관광지의 요인별 감성 점수를 평균을 취해서 실제로 관광객의 방문이 많은 관광지의 감성 점수가 과소평가되었을 수 있다. 관광지별로 정확한 외국인 방문객 수에 대한 데이터가 있다면 더 정확한 외국인 관광객 수에 영향을 끼치는 주요 요인을 추출할 수 있을 것이다.

참고문헌

Bae, W. S., Kim, M. J., and Kim, C. G. (2014), The General Linear Test in the Ridge Regression, *Communications for Statistical Applications and Methods*, **21**(4), 297-307.

Cho, S., Kim, B., Park, M., Lee, G., and Kang, P. (2017), Extraction of Satisfaction Factors and Evaluation of Tourist Attractions based on Travel Site Review Comments, *Journal of Korean Institute of Industrial Engineers*, **43**(1), 62-71.

Choi, A., Wang, S., and Koo, H. G. (2018), A comparative Study on Satisfaction and Influence Factors of Chinese, Japanese and American tourists in Korea, *Journal of Digital Convergence*, **16**(5), 123-135.

Chun, M. H. (2011), Credibility of e-WOM in Travel Industry, and Its Influence in WOM Effect, *The Journal of the Korea Contents Association*, **11**(5), 424-432.

Cule, E., Vineis, P., and De Iorio, M. (2011), Significance Testing in Ridge Regression for Genetic Data, *BMC Bioinformatics*, **12**(1), 372.

Jiang, W. (2015), A Study on the Satisfaction Factors for Foreign Tourism Services, Kookmin University.

Khatibi, A., Belém, F., da Silva, A. P. C., Almeida, J. M., and Gonçalves, M. A. (2020), Fine-Grained Tourism Prediction : Impact of Social and Environmental Features, *Information Processing & Management*, **57**(2), 102057.

Kim, J. H. (1995), A Study on Reliability Analysis of Questionnaire Items, Jeonju University.

Kim, S. B., Kwon, S. J., and Kim, J. T. (2015), Building Sentiment Dictionary and Polarity Classification of Blog Review By Using Elastic Net, In *Proceedings of the 2015 Winter Conference of Korean Institute of Information Scientists and Engineers*.

Lee, D. D. and Seung, H. S. (1999), Learning the Parts of Objects by Non-Negative Matrix Factorization, *Nature*, **401**(6755), 788-791.

Lee, J. Y. and Tsou, M.-H. (2018), Mapping Spatiotemporal Tourist Behaviors and Hotspots Through Location-Based Photo-Sharing Service (Flickr) Data BT-Progress in Location Based Services 2018, in Kiefer, P. et al. (eds). Cham : Springer International Publishing, 315-334.

Levy, O. and Goldberg, Y. (2014), Neural Word Embedding as Implicit Matrix Factorization, *In Advances in Neural Information Processing Systems*, 2177-2185.

Lim, H. S. and Nam, Y. S. (2017), A Research on Effects of the Satisfaction and Revisiting Intension by Image of Tourist Destination : Focus on Chinese Tourist of Seongsan Sunrise Peak, *The Journal of the Korea Contents Association*, **17**(2), 298-307.

Önder, I., Koerbitz, W., and Hubmann-Haidvogel, A. (2014), Tracing Tourists by Their Digital Footprints : The Case of Austria, *Journal of Travel Research*, **55**(5), 566-573.

Pang, B., Lee, L., and Vaithyanathan, S. (2002), Thumbs up? Sentiment Classification Using Machine Learning Techniques, *arXiv preprint cs/0205070*.

Seoul Metropolitan Government (2018), Seoul Tourism Plan 2023.
Shahnaz et al. (2006), Document Clustering Using Nonnegative Matrix Factorization, *Information Processing & Management*, **42**(2), 373-386.
Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018), Short-Text Topic Modeling Via Non-Negative Matrix Factorization Enriched with Local Word-Context Correlations, *In Proceedings of the 2018 World Wide Web Conferenc*, 1105-1114.
Yang, X., Pan, B., Evans, J. A., and Lv, B. (2015), Forecasting Chinese Tourist Volume with Search Engine Data, *Tourism Management*, **46**, 386-397.
Yeum S. K. and Lee J. S. (2015), A Study on The Relationship Between City Tour Components, City Image and Tourist Satisfaction 1 : Focused on Seoul City Tour, *Journal of Tourism and Leisure Research*, **27**(2), 171-186.
Yu, Y.-H., Lee, J.-Y., and Lee, H.-C. (2016), Determinants of Revisit Intention to Seoul by Chinese Tourists : The Comparison of Group Tourist and FIT, *Journal of Tourism and Leisure Research*, **28**(8), 241-256.
Zou, H. and Hastie, T. (2005), Regularization and Variable Selection Via the Elastic Net, *Journal of the Royal Statistical Society : Series B (statistical methodology)*, **67**(2), 301-320.

저자소개

박홍제 : 서울과학기술대학교 산업공학과에서 2018년 학사학위를 취득하고 서울과학기술대학교에서 데이터사이언스학과 석사과정에 재학 중이다. 연구 분야는 데이터마이닝, 기계학습 등이다.

이도현 : 한남대학교 글로벌비지니스학과에서 2018년 학사학위를 취득하였다. 그 후 서울과학기술대학교 데이터사이언스학과에서 2020년 8월 석사학위를 취득 후, 현재 동일 학과에서 박사과정에 재학 중이다. 연구 분야는 데이터마이닝, 기계학습 등이다.

김경옥 : 김경옥 교수는 POSTECH 신소재공학과에서 2008 학사, POSTECH 산업경영공학과에서 2013년 박사학위를 취득하였다. 삼성경제연구소 연구원을 거쳐 2015년부터 서울과학기술대학교 산업공학과에서 조교수로 재직하고 있다. 연구 분야는 데이터마이닝, 기계학습 등이다.